Stability of Cluster Analysis

Matthias Templ[1,2] and Peter Filzmoser[2]

[1] Statistics Austria, Vienna, Austria
[2] Department of Statistics and Probability Theory, Vienna University of Technology

Abstract

Cluster analysis is a method for finding groups in multivariate data without providing any information about group membership (unsupervised classification). Many different clustering algorithms have been proposed in the literature, and many methods are implemented in R. Unfortunately, for real data sets without obvious grouping structure, different cluster algorithms will in general give slightly different results, sometimes even completely different results. For the user it would thus be important to know which clustering methods are "ideally suited" for analyzing the data at hand.

To start with, one first has to think about preparing the data for clustering. Is it necessary to transform or to scale the data?

Secondly, some cluster algorithms are based on distances. Which distance measure is appropriate? Will the results heavily depend on the distance measure used?

Thirdly, which clustering method should be chosen? Some cluster methods require knowledge on the number of clusters. Will the method and/or the selected number of clusters heavily influence the outcome? Which clustering methods give stable results?

A number of different validity measures have been proposed which help to determine the "correct" number of clusters. Which validity measures are really helpful with this decision?

We will try to provide answers to the above questions using a real data set from geochemistry. A tool in R has been developed which allows a flexible handling of various clustering methods based on different distance measures and evaluated on different validity measures.

Keywords: Cluster analysis, Multivariate methods, Stability, Robustness